Supplementary material: Riemannian stochastic quasi-Newton algorithm with variance reduction and its convergence analysis

A Manifolds in numerical comparison

This section gives a brief explanation of the manifolds that appear in the numerical comparisons in Section 5.

A.1 SPD manifold

Let $\mathcal{S}^d_{++}$ be the manifold of $d \times d$ SPD matrices [35]. If we endow $\mathcal{S}^d_{++}$ with the Riemannian metric [36] defined by

$$\langle \xi_{\mathbf{X}}, \eta_{\mathbf{X}} \rangle_{\mathbf{X}} \quad = \quad \mathrm{trace}(\xi_{\mathbf{X}} \mathbf{X}^{-1} \eta_{\mathbf{X}} \mathbf{X}^{-1}), \quad \xi_{\mathbf{X}}, \eta_{\mathbf{X}} \in T_{\mathbf{X}} \mathcal{S}^d_{++},$$

at $\mathbf{X} \in \mathcal{S}^d_{++}$, the SPD manifold $\mathcal{S}^d_{++}$ becomes a Riemannian manifold. The explicit formula for the exponential mapping with repsect to this metric is given by

$$\mathrm{Exp}_{\mathbf{X}}(\xi_{\mathbf{X}}) \quad = \quad \mathbf{X}^{1/2} \exp(\mathbf{X}^{-1/2} \xi_{\mathbf{X}} \mathbf{X}^{-1/2}) \mathbf{X}^{1/2}$$

for any $\xi_{\mathbf{X}} \in T_{\mathbf{X}} \mathcal{S}^d_{++}$ and $\mathbf{X} \in \mathcal{S}^d_{++}$, where $\exp(\cdot)$ is the matrix exponential function. On the other hand, $R_{\mathbf{X}}(\xi_{\mathbf{X}}) = \mathbf{X} + \xi_{\mathbf{X}} + \frac{1}{2} \xi_{\mathbf{X}} \mathbf{X}^{-1} \xi_{\mathbf{X}}$ proposed in [37] is a retraction, which is symmetric positive-definite for all $\xi_{\mathbf{X}} \in T_{\mathbf{X}} \mathcal{S}^d_{++}$ and $\mathbf{X} \in \mathcal{S}^d_{++}$. The parallel translation on $\mathcal{S}^d_{++}$ along $\eta_{\mathbf{X}}$ is given by

$$P_{\eta_{\mathbf{X}}}(\xi_{\mathbf{X}}) \quad = \quad \mathbf{X}^{1/2} \mathbf{Y} \mathbf{X}^{-1/2} \xi_{\mathbf{X}} \mathbf{X}^{-1/2} \mathbf{Y} \mathbf{X}^{1/2},$$

where $\mathbf{Y} = \exp(\mathbf{X}^{-1/2} \eta_{\mathbf{X}} \mathbf{X}^{-1/2}/2)$. A more efficient algorithm that constructs an isometric vector transport is proposed based on a field of orthonormal tangent bases [31] while satisfying the locking condition in Assumption 3. We use it in the experiment, and the details are in [21, 31]. The logarithm map of $\mathbf{Y}$ at $\mathbf{X}$ is given by

$$\mathrm{Log}_{\mathbf{X}}(\mathbf{Y}) \quad = \quad \mathbf{X}^{1/2} \log(\mathbf{X}^{-1/2} \mathbf{Y} \mathbf{X}^{-1/2}) \mathbf{X}^{1/2} = \log(\mathbf{Y} \mathbf{X}^{-1}) \mathbf{X},$$

where $\log(\cdot)$ is the matrix logarithm function.

A.2 Grassmann manifold

A point on the Grassmann manifold is an equivalence class represented by a $d \times r$ orthogonal matrix $\mathbf{U}$ with orthonormal columns, i.e., $\mathbf{U}^T \mathbf{U} = \mathbf{I}$. Two $d \times r$ orthogonal matrices express the same element on the Grassmann manifold if they are mapped to each other by the right multiplication of an $r \times r$ orthogonal matrix $\mathbf{O} \in \mathcal{O}(r)$. Equivalently, an element of $\mathrm{Gr}(r, d)$ is identified with a set of $d \times r$ orthogonal matrices $[\mathbf{U}] := \{\mathbf{U}\mathbf{O} : \mathbf{O} \in \mathcal{O}(r)\}$. That is, $\mathrm{Gr}(r, d) := \mathrm{St}(r, d)/\mathcal{O}(r)$, where $\mathrm{St}(r, d)$ is the *Stiefel manifold*, which is the set of matrices of size $d \times r$ with orthonormal columns. The Grassmann manifold has the structure of a Riemannian quotient manifold [1, Section 3.4].

The exponential mapping for the Grassmann manifold from $\mathbf{U}(0) := \mathbf{U} \in \mathrm{Gr}(r, d)$ in the direction of $\xi \in T_{\mathbf{U}(0)} \mathrm{Gr}(r, d)$ is given in a closed form as [38]

$$\mathbf{U}(t) \quad = \quad [\mathbf{U}(0)\mathbf{V} \ \ \mathbf{W}] \begin{bmatrix} \cos t\Sigma \\ \sin t\Sigma \end{bmatrix} \mathbf{V}^T,$$

where $\xi = \mathbf{W} \Sigma \mathbf{V}^T$ is the singular value decomposition (SVD) of ξ with rank r. The $\sin(\cdot)$ and $\cos(\cdot)$ operations are performed only on the diagonal entries. The parallel translation of $\zeta \in T_{\mathbf{U}(0)} \mathrm{Gr}(r, d)$ on the Grassmann manifold along $\gamma(t)$ with $\dot{\gamma}(0) = \mathbf{W} \Sigma \mathbf{V}^T$ is given in a closed form by

$$\zeta(t) \quad = \quad \left([\mathbf{U}(0)\mathbf{V} \ \ \mathbf{W}] \begin{bmatrix} -\sin t\Sigma \\ \cos t\Sigma \end{bmatrix} \mathbf{W}^T + (\mathbf{I} - \mathbf{W}\mathbf{W}^T) \right) \zeta.$$

The logarithm map of $\mathbf{U}(t)$ at $\mathbf{U}(0)$ on the Grassmann manifold is given by

$$\mathrm{Log}_{\mathbf{U}(0)}(\mathbf{U}(t)) \;=\; \mathbf{W}\arctan(\Sigma)\mathbf{V}^T,$$

where $\mathbf{W}\Sigma\mathbf{V}^T$ is the SVD of $(\mathbf{U}(t) - \mathbf{U}(0)\mathbf{U}(0)^T\mathbf{U}(t))(\mathbf{U}(0)^T\mathbf{U}(t))^{-1}$ with rank r. Furthermore, a popular retraction is

$$R_{\mathbf{U}(0)}(\xi) = \mathrm{qf}(\mathbf{U}(0) + t\xi) \quad (= \mathbf{U}(t)),$$

which extracts the orthonormal factor based on the QR decomposition, and a popular vector transport uses an orthogonal projection of ξ to the horizontal space at $\mathbf{U}(t)$, i.e., $(\mathbf{I} - \mathbf{U}(t)\mathbf{U}(t)^T)\xi$ [1].

B Two-loop Hessian inverse update algorithm

This section summarizes the Riemannian two-loop Hessian inverse updating algorithm in Algorithm A.1. This is a straightforward extension of that in the Euclidean space explained in [27, Section 7.2].

Algorithm A.1 Hessian inverse update

Require: Memory depth τ, correction pairs $\{s_u^k, y_u^k\}_{u=k-\tau}^{k-1}$, gradient p.

```
p_0 = p.
H_k^0 = χ_k id = ⟨s_{k-1}^k, y_{k-1}^k⟩ / ⟨y_{k-1}^k, y_{k-1}^k⟩ id.
for u = 0, 1, 2, ..., τ − 1 do
    ρ_{k-u} = 1/⟨s_{k-u-1}^k, y_{k-u-1}^k⟩.
    α_u = ρ_{k-u-1}⟨s_{k-u-1}^k, p_u⟩.
    p_{u+1} = p_u − α_u y_{k-u-1}^k.
end for
q_0 = H_k^0 p_τ.
for u = 0, 1, 2, ..., τ − 1 do
    β_u = ρ_{k-τ+u}⟨y_{k-τ+u}^k, q_u⟩.
    q_{u+1} = q_u + (α_{τ-u-1} − β_u) s_{k-τ+u}^k.
end for
q = q_τ.
```

C Proofs of convergence analysis on non-convex functions

This section presents proofs of the global convergence analysis on non-convex functions. In this supplement, only sketches of some proofs are provided, or some proofs are omitted. Hereinafter, we use $\mathbb{E}[\cdot]$ to express expectation with respect to the joint distribution of all random variables. For example, w_t $(= w_t^k)$ is determined by the realizations of the independent random variables $\{i_0^0, i_1^0, \ldots, i_{m_0-1}^0, \ldots, i_0^k, i_1^k, \ldots, i_{t-1}^k\}$, and the total expectation of $f(w_t)$ for any $t \in \mathbb{N}$ can be taken as $\mathbb{E}[f(w_t)] = \mathbb{E}_{i_0^0}\mathbb{E}_{i_1^0}\cdots\mathbb{E}_{i_{t-1}^k}[f(w_t)]$. We also use $\mathbb{E}_{i_t}[\cdot]$ to denote an expected value taken with respect to the distribution of the random variable i_t. Moreover, we omit the subscript $\tilde{w}^k$ for a Riemannian metric $\langle\cdot,\cdot\rangle_{\tilde{w}^k}$ when the tangent space to be considered is clear.

C.1 Eigenvalue bounds of $\mathcal{H}_t^k$ on non-convex functions

We present an essential proposition that bounds the eigenvalues of $\mathcal{H}_t^k$ at w_t, i.e., $\mathcal{H}_t^k := \mathcal{T}_{\tilde{w}^k}^{w_t} \circ \tilde{\mathcal{H}}^k \circ (\mathcal{T}_{\tilde{w}^k}^{w_t})^{-1}$. To this end, we use the *Hessian approximation operator* $\tilde{\mathcal{B}}^k = (\tilde{\mathcal{H}}^k)^{-1}$ instead of $\tilde{\mathcal{H}}^k$. As mentioned in the algorithm description, we consider curvature information for $\tilde{\mathcal{H}}^k$ at $\tilde{w}^k$, i.e., every outer epoch, and reuse this $\tilde{\mathcal{H}}^k$ in the calculation of the second-order modified stochastic gradient $\mathcal{H}_t^k\xi_t$ at w_t. Thus, the proof consists of two steps as follows:

1. We first address the bounds of $\tilde{\mathcal{H}}^k$ at $\tilde{w}^k$. The main task of the proof is to bound the Hessian operator $\tilde{\mathcal{B}}^k = (\tilde{\mathcal{H}}^k)^{-1}$.

2. We bound $\mathcal{H}_t^k$ at w_t based on the bounds of $\tilde{\mathcal{H}}^k$ at $\tilde{w}^k$.

It should be noted that in this subsection, the curvature pair $\{s_j^k, y_j^k\}_{j=k-L}^{k-1} \in T_{\tilde{w}^k}\mathcal{M}$ is simply denoted by $\{s_j, y_j\}_{j=k-L}^{k-1}$.

We first state a lemma for the bound of $\frac{\langle y_k, y_k \rangle}{\langle s_k, y_k \rangle}$.

Lemma C.1. *Suppose Assumption 1 holds. There exists a constant $\Upsilon_{nc} > 0$ such that for all k*

$$\frac{\langle y_k, y_k \rangle}{\langle s_k, y_k \rangle} \quad \leq \quad \Upsilon_{nc}.$$

Proof. We directly obtain $\frac{\|s_k\|^2}{\langle y_k, s_k \rangle} \leq \frac{1}{\epsilon}$ from (4) in the cautious update. Then, we obtain the upper bound of $\frac{\langle y_k, y_k \rangle}{\langle s_k, y_k \rangle}$ as below taking also into account the fact $\|y_k\| \leq c_1 \|s_k\|$ (Lemma 3.9 in [21]), where $c_1 > 0$ is a contant,

$$\frac{\langle y_k, y_k \rangle}{\langle s_k, y_k \rangle} \quad = \quad \frac{\|s_k\|^2}{\langle s_k, y_k \rangle} \cdot \frac{\|y_k\|^2}{\|s_k\|^2} \quad \leq \quad \frac{c_1^2}{\epsilon} \quad (= \Upsilon_{nc}).$$

Denoting c_1^2/ϵ as Υ_{nc}, this completes the proof. ☐

Next, we bound trace($\hat{\tilde{\mathcal{B}}}$) to bound the eigenvalues of $\tilde{\mathcal{H}}^k$, where a *hat* denotes the coordinate expression of the operator. The basic structure of the proof follows those of stochastic L-BFGS methods in the Euclidean space, e.g., [25, 16, 15]. Nevertheless, some special treatment is required in light of the Riemannian setting. It should be noted that trace($\hat{\tilde{\mathcal{B}}}$) does not depend on the chosen basis.

Lemma C.2 (Bounds of trace of $\tilde{\mathcal{B}}^k$)**.** *Consider the recursion of $\tilde{\mathcal{B}}_u^k$ as*

$$\tilde{\mathcal{B}}_{u+1}^k \quad = \quad \check{\mathcal{B}}_u^k - \frac{\check{\mathcal{B}}_u^k s_{k-\tau+u} (\check{\mathcal{B}}_u^k s_{k-\tau+u})^\flat}{(\check{\mathcal{B}}_u^k s_{k-\tau+u})^\flat s_{k-\tau+u}} + \frac{y_{k-\tau+t} y_{k-\tau+u}^\flat}{y_{k-\tau+u}^\flat s_{k-\tau+u}}, \tag{A.1}$$

where $\check{\mathcal{B}}_u^k = \mathcal{T}_{\tilde{w}^{k-1}}^{\tilde{w}^k} \circ \tilde{\mathcal{B}}_u^k \circ (\mathcal{T}_{\tilde{w}^{k-1}}^{\tilde{w}^k})^{-1}$ for $u = 0, \ldots, \tau - 1$. The Hessian approximation at the k-th outer epoch is $\tilde{\mathcal{B}}^k = \tilde{\mathcal{B}}_\tau^k$ when $u = \tau - 1$. Then, consider the Hessian approximation $\tilde{\mathcal{B}}^k = \tilde{\mathcal{B}}_\tau^k$ in (A.1) with $\tilde{\mathcal{B}}_0^k = \frac{\langle y_k, y_k \rangle}{\langle s_k, y_k \rangle}\mathrm{id}$. If Assumption 1 holds, trace($\hat{\tilde{\mathcal{B}}}^k$) *in a coordinate expression of $\tilde{\mathcal{B}}^k$ is uniformly upper bounded for all $k \geq 1$ as*

$$\mathrm{trace}(\hat{\tilde{\mathcal{B}}}^k) \quad \leq \quad (M + \tau)\Upsilon_{nc},$$

where M is the dimension of $\mathcal{M}$. Here, a hat expression represents the coordinate expression of an operator.

Proof. The bound of trace($\hat{\tilde{\mathcal{B}}}^k$) is first obtained from Lemma C.1. Then, calculating recursively the obtained relation, we can bound trace($\hat{\tilde{\mathcal{B}}}^k$) by the initial value of trace($\hat{\tilde{\mathcal{B}}}^k$). Finally, bounding the initial value of trace($\hat{\tilde{\mathcal{B}}}^k$) by $M\Upsilon_{nc}$, we obtain the claim. Since the proof can be completed in parallel to the Euclidean case [17] and the Riemannian case [29], we omit the complete proof. ☐

We further provide the bounds of $\tilde{\mathcal{H}}^k$.

Lemma C.3 (Bounds of $\tilde{\mathcal{H}}^k$ on non-convex functions)**.** *If Assumption 1 holds, the eigenvalues of $\tilde{\mathcal{H}}^k$ is bounded by some positive constants γ_{nc} and Γ_{nc} for all $k \geq 1$ as*

$$\gamma_{nc}\mathrm{id} \ \preceq \ \tilde{\mathcal{H}}^k \ \preceq \ \Gamma_{nc}\mathrm{id}.$$

Proof. The proof first provides the bound of the sum of the eigenvalues of $\hat{\tilde{\mathcal{B}}}^k$ from Lemma C.2. Then, the bound of $\tilde{\mathcal{H}}^k$ is given. The proof for the lower bound is obtained in parallel to the Euclidean case [25]. Moreover, the upper bound is given by extending the proof of [15]. We omit the complete proof. ☐

Finally, we present the proposition for the bounds of $\mathcal{H}_t^k$ on non-convex functions.

Proposition C.4 (Bounds of $\mathcal{H}_t^k$ on non-convex functions)**.** *Consider the operator* $\mathcal{H}_t^k := \mathcal{T}_{\tilde{w}^k}^{w_t} \circ \tilde{\mathcal{H}}^k \circ (\mathcal{T}_{\tilde{w}^k}^{w_t})^{-1}$*. If Assumption 1 holds, the range of eigenvalues of* $\mathcal{H}_t^k$ *is bounded below by* γ_{nc} *and above by* Γ_{nc} *for all* $k \geq 1, t \geq 1$*, i.e.,*

$$\gamma_{nc}\mathrm{id} \ \preceq \ \mathcal{H}_t^k \ \preceq \ \Gamma_{nc}\mathrm{id}, \tag{A.2}$$

where γ_{nc} *and* Γ_{nc} *are some positive constants.*

Proof. Noting that $\mathcal{H}_t^k := \mathcal{T}_{\tilde{w}^k}^{w_t} \circ \tilde{\mathcal{H}}^k \circ (\mathcal{T}_{\tilde{w}^k}^{w_t})^{-1}$, where $\tilde{\eta}_t = R_{\tilde{w}^k}^{-1}(w_t)$, and that $\mathcal{T}_{\tilde{w}^k}^{w_t}$ is a linear transformation operator, we can conclude that the eigenvalues of $\mathcal{H}_t^k$ and $\tilde{\mathcal{H}}^k$ are identical. In fact, let hat expressions be representation matrices with some bases of $T_{w_t}\mathcal{M}$ and $T_{\tilde{w}^k}\mathcal{M}$. We then have the relation $\det(\mu\mathbf{I} - \hat{\mathcal{H}}_t^k) = \det(\mu\mathbf{I} - \hat{\tilde{\mathcal{H}}}^k)$. Consequently, Lemma C.3 directly yields the claim. This completes the proof. ☐

C.2 Proof of global convergence analysis (Theorem 4.1)

Proof. The proof is provided by extending that of [22] with careful treatment of $\mathcal{H}_t^k$. We also refer to that of [39] in the Euclidean space. We omit the complete proof. ☐

C.3 Proof of global convergence rate analysis (Theorem 4.2)

The global convergence rate analysis on non-convex functions in the Euclidean SVRG has been proposed in [12]. Its further extensions to the stochastic L-BFGS setting and the Riemannian setting have been proposed in [15] and [23], respectively. The proof in this subsection mainly follows that in [12] by integrating its two extensions in [15, 23]. Moreover, retraction and vector transport are carefully treated in the proof. Finally, it should be noted that, since this section discusses the k-th epoch, we omit the superscript 'k'. Moreover, we also omit the subscript w_t for a Riemannian metric $\langle \cdot, \cdot \rangle_{w_t}$ when its point is apparent.

C.3.1 Essential propositions

This subsection first presents an essential lemma concerning the bound of $\mathbb{E}_{i_t}[\|\xi_t\|^2]$, where the vector transport is carefully handled. Proposition C.6 is then presented by extending [12, 15, 23]. It should be noted that we carefully treat the difference between the exponential mapping and retraction for Proposition C.6.

We first present an essential lemma.

Lemma C.5. *Suppose Assumption 1, which guarantees Lemmas 3.9, 3.10, and 3.11 for* $\bar{w} = w^*$*. Let* $L_l > 0$ *be a constant such that*

$$\|P(\gamma)_z^w(\mathrm{grad} f_i(z)) - \mathrm{grad} f_i(w)\|_w \leq L_l \mathrm{dist}(z, w), \qquad w, z \in \Theta, \ i = 1, 2, \ldots, n.$$

The existence of such an L_l *is guaranteed by Lemma 3.11. Then, the upper bound of the variance of* $\mathbb{E}_{i_t}[\|\xi_t\|^2]$ *is given by*

$$\mathbb{E}_{i_t}[\|\xi_t\|^2] \quad \leq \quad 4(L_l^2 + \tau_2^2 C^2 \theta^2)(\mathrm{dist}(w_t, \tilde{w}))^2 + 2\|\mathrm{grad} f(w_t)\|^2.$$

Proof. The proof is similar to that of Lemma 5.8 in [22]. We omit the detail of the proof. ☐

Proposition C.6. *Let* $\mathcal{M}$ *be a Riemannian manifold and* $w^* \in \mathcal{M}$ *be a non-degenerate local minimizer of* f *(i.e.,* $\mathrm{grad} f(w^*) = 0$*, and the Hessian* $\mathrm{Hess} f(w^*)$ *of* f *at* w^* *is positive definite). Suppose Assumption 1 holds. Let the constants* θ *be in (5),* τ_1 *and* τ_2 *be in (6),* L_l *be in (7),* γ_{nc} *and* Γ_{nc} *be in (8), and* L *be in Lemma 3.4. For* $c_t, c_{t+1}, \nu_t > 0$*, we set*

$$c_t \quad = \quad c_{t+1}\left(1 + \alpha_t \nu_t + 4\zeta\alpha_t^2(L_l^2 + \tau_2^2 C^2\theta^2)\frac{\Gamma_{nc}^2}{\tau_1^2}\right) + 2\alpha_t^2 L(L_l^2 + \tau_2^2 C^2\theta^2)\Gamma_{nc}^2. \tag{A.3}$$

We also define

$$\Delta_t \quad := \quad \alpha_t\left(\gamma_{nc} - \frac{c_{t+1}\Gamma_{nc}^2}{\nu_t \tau_1^2} - \alpha_t L \Gamma_{nc}^2 - 2c_{t+1}\zeta\alpha_t\frac{\Gamma_{nc}^2}{\tau_1^2}\right). \tag{A.4}$$

Let α_t, ν_t, and c_{t+1} be defined such that it holds $\Delta_t > 0$. It then follows that for any sequence $\{\tilde{w}_t\}$ generated by Algorithm 1 with Option I-B and with a fixed step size $\alpha_t := \alpha$ and $m_k := m$ converging to w^*, the expected squared norm of the Riemannian gradient, $\mathrm{grad} f(w_t)$*, satisfies the bound as*

$$\mathbb{E}[\|\mathrm{grad} f(w_t)\|^2] \quad \leq \quad \frac{V_t - V_{t+1}}{\Delta_t}, \tag{A.5}$$

where $V_t := \mathbb{E}[f(w_t) + c_t(\mathrm{dist}(\tilde{w}, w_t))^2]$ for $0 \leq k \leq K-1$.

Proof. The sketch of the proof is as follows: We first obtain the relation between $\mathbb{E}[f(w_{t+1})]$ and $f(w_t)$ from Lemma 3.4. We then bound the expected squared distance between $\tilde{w}$ and w_{t+1}, i.e., $\mathbb{E}[(\mathrm{dist}(\tilde{w}, w_{t+1}))^2]$, from Lemma 6 in [40] by considering (6) in Lemma 3.10 and Proposition C.4. We also use $\mathbb{E}_{i_t}[\mathcal{H}_t \xi_t] = \mathcal{H}_t \mathrm{grad} f(w_t)$. Next, we introduce a function defined as $V_t := \mathbb{E}[f(w_t) + c_t(\mathrm{dist}(\tilde{w}, w_t))^2]$, which measures how far the given parameter w_t is from $\tilde{w}$ and the objective function value. Finally, calculating V_{t+1} from Lemma C.5, we obtain the claim. □

The following proposition is very similar to Theorem 2 in [12].

Proposition C.7 (Theorem 2 in [12])**.** *Let $\mathcal{M}$ be a Riemannian manifold and $w^* \in \mathcal{M}$ be a non-degenerate local minimizer of f. Consider Algorithm 1 with Option I-B and II-A, and suppose Assumption 1 holds. Let the constants θ be in (5), τ_1 and τ_2 be in (6), and L_l be in (7). γ_{nc} and Γ_{nc} are the constants in (8). Let $c_m = 0$, $\alpha_t = \alpha > 0, \nu_t = \nu > 0$, and c_t is defined as (A.3) such that Δ_t defined in (A.4) satisfies $\Delta_t > 0$ for $0 \leq t \leq m-1$. Define $\delta_t := \min_t \Delta_t$. Let T be mK. It then follows that for the output w_{sol} of Algorithm 1,*

$$\mathbb{E}[\|\mathrm{grad} f(w_{\mathrm{sol}})\|^2_{w_{\mathrm{sol}}}] \quad \leq \quad \frac{f(w^0) - f(w^*)}{T\delta_t}. \tag{A.6}$$

Proof. Because the proof is identical to those in [12, 15, 23], we omit the detail. The complete proof is there. A sketch of the proof is as follows: We first telescope the sum of (A.5) from $t = 0$ to $t = m-1$ by introducing δ_t, and estimate its upper bound from the difference between V_0^s and V_0^m. After showing that this difference is equivalent to the expected difference between $f(\tilde{w}^k)$ and $f(\tilde{w}^{k+1})$, summing up from $k = 0$ to $k = K-1$, we obtain the desired claim. □

C.3.2 Main proof of Theorem 4.2

Proof. The proof is based on the extensions of results in [12, 15, 23]. We omit the complete proof. A sketch of the proof is as follows: From (A.4) in Proposition C.6, we need to consider the upper bound of c_t defined in (A.3). To this end, the upper bound of c_0 is first derived. For this particular purpose, denoting, for simplicity, $\varphi = \alpha\nu + 4\zeta\alpha^2(L_l^2 + \tau_2^2 C^2\theta^2)\frac{\Gamma_{nc}^2}{\tau_1^2}$ and $\omega = \tau_2 C\theta$, we first give the bound of φ as $\varphi \in \left(\frac{\mu_0\zeta^{1-2a_2}}{n^{3a_1/2}}, 5\frac{\mu_0\zeta^{1-2a_2}}{n^{3a_1/2}}\right)$. Then, considering the recurrence relation $c_t = c_{t+1}(1+\varphi) + 2\alpha^2 L(L_l^2 + \omega^2)\Gamma_{nc}^2$, we obtain the bound of c_0. Next, we attempt to estimate the lower bound of δ_t, i.e., $\min_t \Delta_t$, where the bound of c_0 is used. Finally, substituting the lower bound of δ_t into (A.6) in Proposition C.7 completes the proof. □

D Proof of local convergence analysis on retraction strongly convex functions

This section presents a local convergence rate analysis in a neighborhood of a local minimum for retraction strongly convex functions. This *local* setting is very common and standard in manifold optimization.

D.1 Eigenvalue bounds of $\mathcal{H}_t^k$ on retraction strongly convex functions

We first bound $\mathrm{trace}(\hat{\tilde{\mathcal{B}}})$ and $\det(\hat{\tilde{\mathcal{B}}})$ to bound the eigenvalues of $\tilde{\mathcal{H}}^k$, where a *hat* denotes the coordinate expression of the operator. The bound of $\mathrm{trace}(\hat{\tilde{\mathcal{B}}})$ is identical to that of the non-convex case in Lemma C.2. Therefore, we concentrate on the bound of $\det(\hat{\tilde{\mathcal{B}}})$. As in Lemma C.2, the proof follows that of stochastic L-BFGS methods in the Euclidean space, e.g., [25, 16, 17]. Similarly to Section C.1, it should be noted that $\mathrm{trace}(\hat{\tilde{\mathcal{B}}})$ and $\det(\hat{\tilde{\mathcal{B}}})$ do not depend on the chosen basis.

Lemma D.1 (Bounds of trace and determinant of $\tilde{\mathcal{B}}^k$)**.** *Consider the recursion of $\tilde{\mathcal{B}}^k_u$ defined in (A.1). If Assumptions 1 and 3 hold,* $\mathrm{trace}(\hat{\tilde{\mathcal{B}}}^k)$ *in a coordinate expression of $\tilde{\mathcal{B}}^k$ is uniformly upper bounded for all $k \geq 1$,*

$$\mathrm{trace}(\hat{\tilde{\mathcal{B}}}^k) \quad \leq \quad (M+\tau)\Upsilon_c,$$

where M is the dimension of $\mathcal{M}$. Similarly, if Assumptions 1 and 3 hold, $\det(\hat{\tilde{\mathcal{B}}}^k)$ *in a coordinate expression of $\tilde{\mathcal{B}}^k$ is uniformly lower bounded for all k as*

$$\det(\hat{\tilde{\mathcal{B}}}^k) \quad \geq \quad \upsilon^M \left[\frac{\mu}{(M+\tau)\Upsilon_c}\right]^\tau.$$

Here, a hat expression represents the coordinate expression of an operator.

Proof. The proof follows that of the Euclidean case [25, 16, 17]. We omit the proof here. □

We next prove a lemma for the bound of $\tilde{\mathcal{H}}^k$.

Lemma D.2 (Bound of $\tilde{\mathcal{H}}^k$ on retraction strongly convex functions)**.** *If Assumptions 1 and 3 hold, the eigenvalues of $\tilde{\mathcal{H}}^k$ are bounded by γ_c and Γ_c with $0 < \gamma_c < \Gamma_c < \infty$ uniformly for all $k \geq 1$ as*

$$\gamma_c \mathrm{id} \ \preceq \ \tilde{\mathcal{H}}^k \ \preceq \ \Gamma_c \mathrm{id}.$$

Proof. The proof is given by exploiting Lemma D.1. The complete proof follows that of the Euclidean case [25, 16, 17] and we omit the detail of it. □

Finally, we give the bounds of $\mathcal{H}^k_t$ on retraction strongly convex functions.

Proposition D.3 (Bounds of $\mathcal{H}^k_t$ for retraction strongly convex functions)**.** *Consider the operator $\check{\mathcal{H}}^k := \mathcal{T}^{w_t}_{\tilde{w}^k} \circ \tilde{\mathcal{H}}^k \circ (\mathcal{T}^{w_t}_{\tilde{w}^k})^{-1}$. If Assumptions 1 and 3 hold, the range of eigenvalues of $\mathcal{H}^k_t$ is bounded by some positive constants γ_c and Γ_c with $\gamma_c < \Gamma_c$ uniformly for all $k \geq 1, t \geq 1$, i.e.,*

$$\gamma_c \mathrm{id} \ \preceq \ \mathcal{H}^k_t \ \preceq \ \Gamma_c \mathrm{id}.$$

Proof. Like Proposition C.4, we can give the proof by exploiting Lemma D.2. Since the proof is identical to that of Proposition C.4, the complete proof is omitted. □

D.2 Proof of local convergence rate analysis (Theorem 4.3)

Proof. The sketch of the proof is as follows: From Lemma 3.4, we first obtain the relation between$f(w_{t+1})$ and $f(w_t)$. Taking expectation of the relation with regard to i_t, we obtain the bound of $\mathbb{E}_{i_t}[f(w_{t+1})] - f(w_t)$ using the fact that $\mathbb{E}_{i_t}[\mathcal{H}^k_t \xi_t] = \mathcal{H}^k_t \mathrm{grad} f(w_t)$ and Proposition D.3. Next, by exploiting the property of the retraction strongly convex, we obtain the new bound of $\mathbb{E}_{i_t}[f(w_{t+1})] - f(w_t)$ with the constant μ of the retraction strongly convex. Plugging the bound of $\mathbb{E}_{i_t}[\|\xi^k_k\|^2]$ (Lemma 5.12 in [22]) into this bound, the bound of $\mathbb{E}_{i_t}[f(w_{t+1})] - f(w_t)$ is furthere obtained. Here, using Lemma 3.5 with $\mathrm{grad} f(w^*) = 0$ and Lemma 3.10, we obtain the lower bounds of $f(w_t) - f(w^*)$ and $f(\tilde{w}^k) - f(w^*)$. Therefore, plugging these into the bound of $\mathbb{E}_{i_t}[f(w_{t+1})] - f(w_t)$, we obtain the new bound. Finally, taking expectations over all random variables and further summing over $t = 0, \ldots, m-1$ of the inner loop on the k-th epoch, we obtain the upper bound of $\mathbb{E}[f(\tilde{w}^{k+1}) - f(w^*)]$. Thus, we obtain the claim. □

E Additional numerical experiments

In this section, we show additional numerical experiments which do not appear in the main text.

E.1 Matrix completion problem on synthetic datasets

E.1.1 Additional results

This section shows the results of six problem instances. We show only the loss on a test set Φ, which is different from the training set Ω. The loss on the test set demonstrates the convergence speed to a satisfactory prediction accuracy of missing entries.

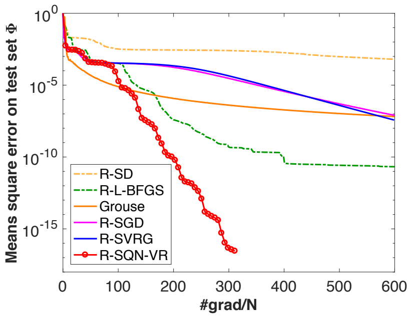
Case MC-S1: We first show the results of the comparison when the number of samples $n = 5000$, the dimension $d = 200$, the memory size $L = 10$, the oversampling ratio (OS) is 8, and the condition number (CN) is 50. We also add Gaussian noise $\sigma = 10^{-10}$. Figures A.1 show the results of four runs excluding the result shown in the main text, which corresponds to "run 1." They show superior performance to other algorithms.

Case MC-S2: influence on low sampling. We look into problem instances from scarcely sampled data, e.g. OS is 4. Other conditions are the same as in **Case MC-S1**. From Figures A.2, we see that the proposed algorithm gives much better and stabler performance against other algorithms.

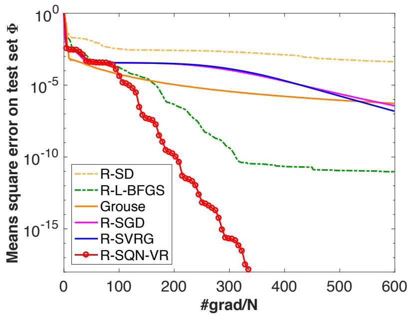
Case MC-S3: influence on ill conditioning. We consider the problem instances with higher condition number (CN) 100. The other conditions are the same as in **Case MC-S1**. Figures A.3 show the superior performances of the proposed algorithm against other algorithms.

Case MC-S4: influence on higher noise. We consider noisy problem instances, where $\sigma = 10^{-6}$. The other conditions are the same as in **Case MC-S1**. Figures A.4 show that the convergent MSE values are much higher than the other cases. Then, we can see the superior performance of the proposed R-SQN-VR against other algorithms.

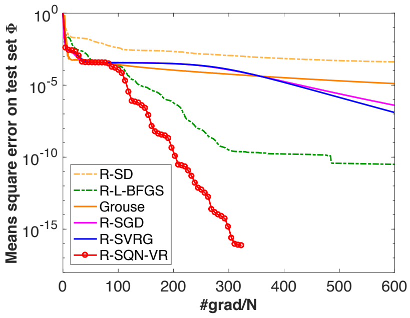
Case MC-S5: influence on higher rank. We consider problem instances with higher rank, where $r = 10$. The other conditions are the same as in **Case MC-S1**. From Figures A.5, the proposed R-SQN-VR still shows superior performance to other algorithms. Grouse indicates a faster decrease in the MSE at the begging of the iterations. However, the convergent MSE values are much higher than those of the other methods.



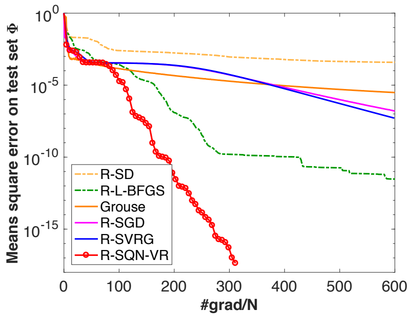
(a) run 2



(b) run 3

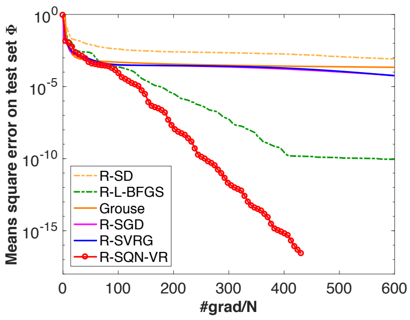


(c) run 4

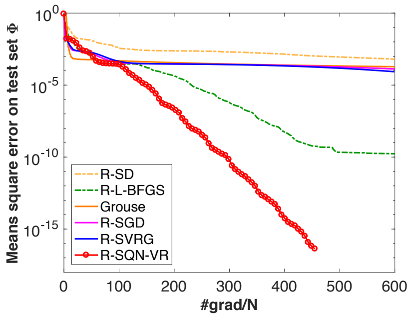


(d) run 5

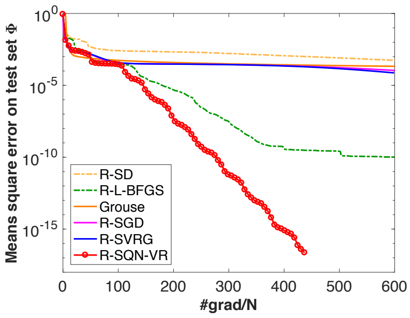
Figure A.1: Performance evaluations on low-rank MC problem (**Case MC-S1: baseline.**).



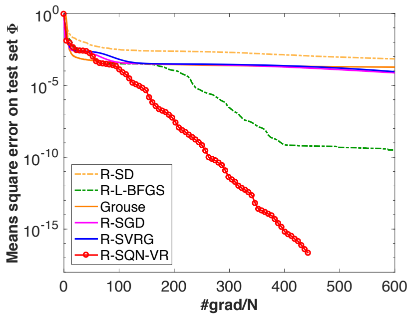
run 2



(b) run 3

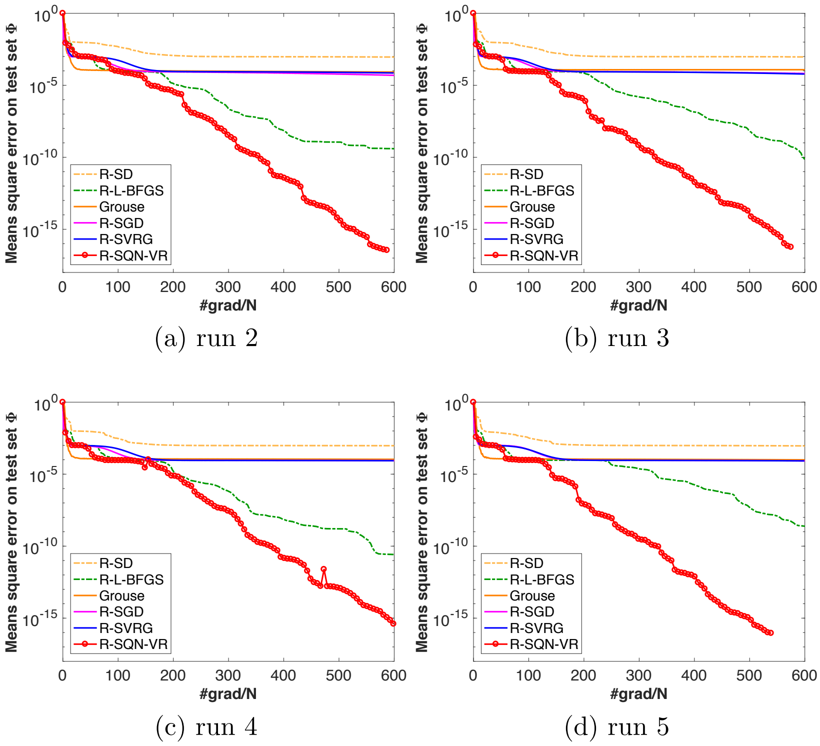


(c) run 4



(d) run 5

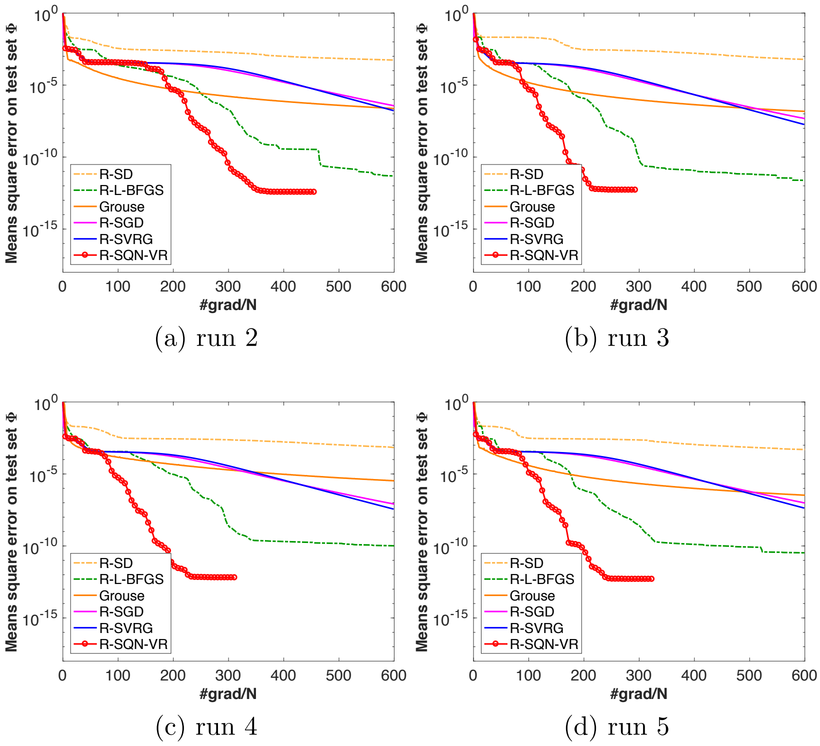
Figure A.2: Performance evaluations on low-rank MC problem (**Case MC-S2: low sampling.**).



(a) run 2 (b) run 3

(c) run 4 (d) run 5

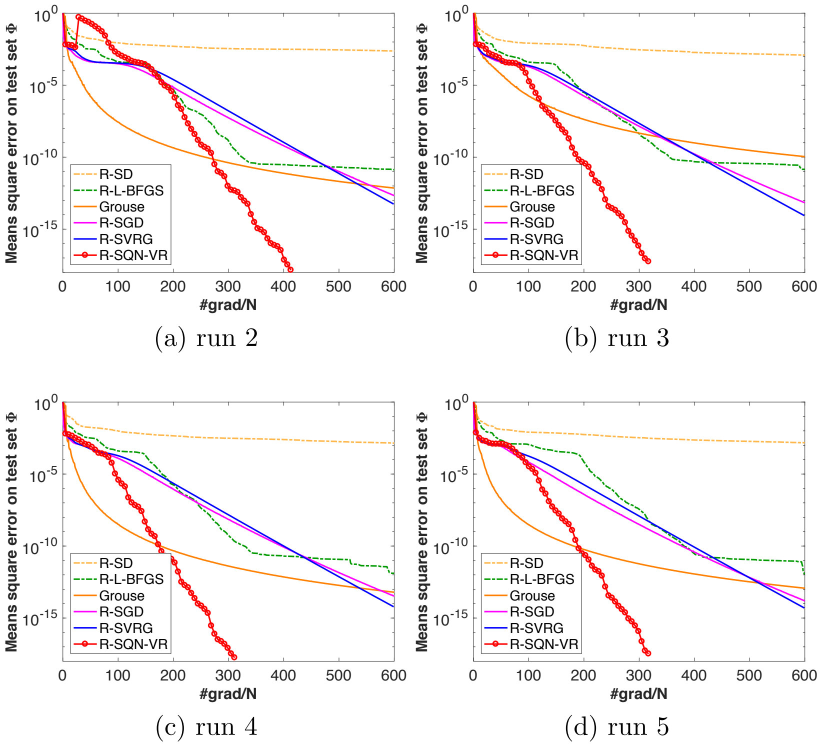
Figure A.3: Performance evaluations on low-rank MC problem (**Case MC-S3: ill-conditioning.**).



(a) run 2 (b) run 3

(c) run 4 (d) run 5

Figure A.4: Performance evaluations on low-rank MC problem (**Case MC-S4: noisy data.**).



(a) run 2

(b) run 3

(c) run 4

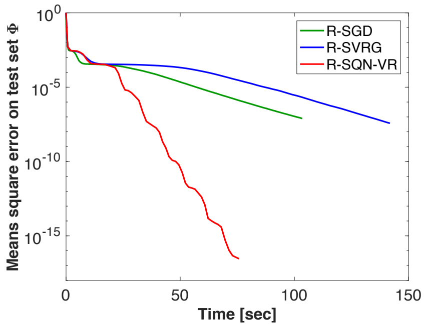
(d) run 5

Figure A.5: Performance evaluations on low-rank MC problem (**Case MC-S5: higher rank.**).

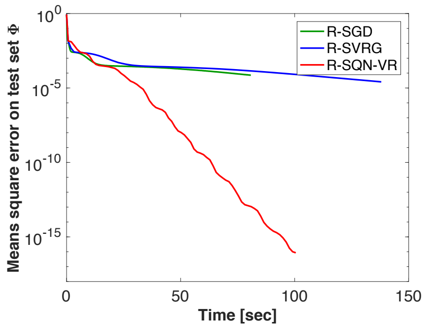
E.1.2 Processing time experiments

The results in terms of the processing time are presented.

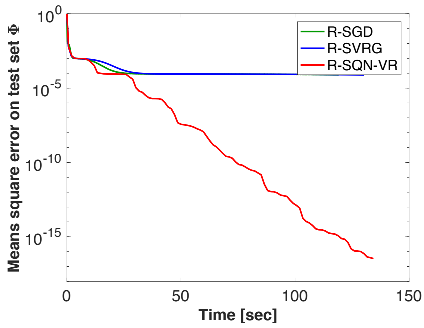
Case MC-S7: Comparison in terms of processing time. Because one major concern of second-order algorithms is, in general, higher computational processing load than first-order algorithms, we additionally show the results in terms of processing times. This evaluation addresses only R-SGD, R-SVRG, and R-SQN-VR because their code structures are similar, whereas the batch-based algorithms, i.e., R-SD and R-L-BFGS, have completely different implementations. Figures A.6 (a)-(e) show the results of the relationship between test MSE and processing time [sec]. From the figures, as expected, R-SGD was much faster in terms of iterations than other algorithms. However, it should be noted that R-SGD suffered from the problem whereby it heavily reduced convergence speed around the solution as reported in the literature. Comparing R-SQN-VR with R-SVRG, R-SQN-VR still yielded better performance, although R-SQN-VR required an additional vector transport of a gradient in each inner iteration and L vector transports of the curvature pairs at every outer epoch than R-SVRG. Overall, R-SQN-VR outperformed R-SGD and R-SVRG in terms of processing time. Consequently, we also confirmed the effectiveness of the proposed R-SQN-VR from the perspective of processing time.



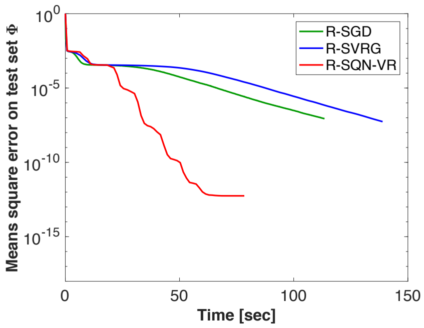
(a) **Case MC-S1: baseline.**



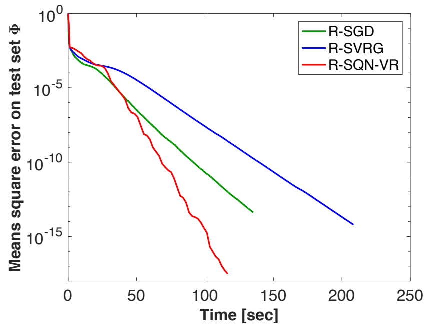
(b) **Case MC-S2: low sampling.**



(c) **Case MC-S3: ill-conditioning.**



(d) **Case MC-S4: noisy data.**



(e) **Case MC-S5: higher rank.**

Figure A.6: Performance evaluations on low-rank MC problem (**Case MC-S7**).

Finally, Figure A.7 shows the results when the memory size of L was changed in R-SQN-VR. Comparing the results with those in Figure 1 (h), cases of smaller sizes improved very slightly, but we did not observe a significant advantage in terms of processing load. From these results in terms of the convergence speed and processing load, we cannot determine the best size of L. This is a subject for future research.

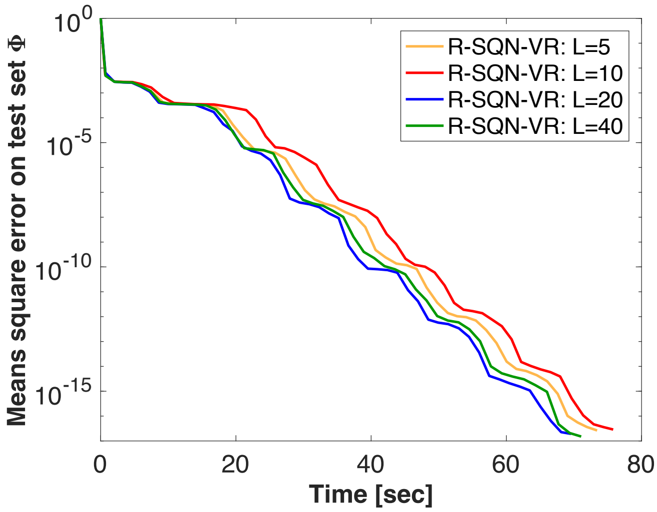
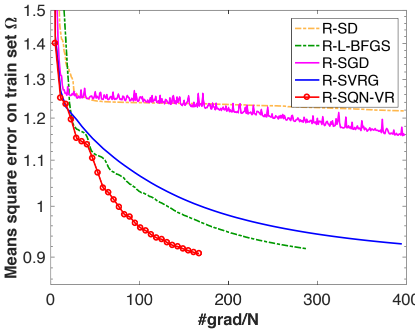


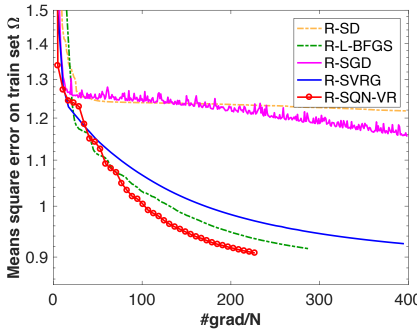
Figure A.7: Performance evaluations on low-rank MC problem (processing time) (**Case MC-S6: different memory sizes**).

E.2 Matrix completion problem on MovieLens 1M dataset

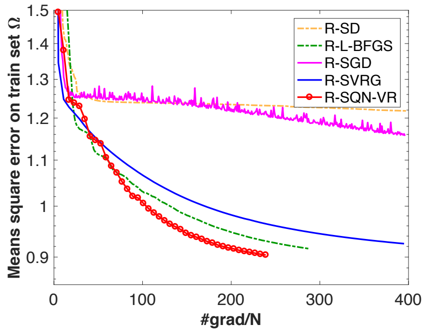
Figures A.8 and A.9 show the results of the cases where $r = 10$ (**MC-R1: lower rank**) and $r = 20$ (**MC-R2: higher rank**), respectively. They show the convergence plots of the training error on Ω and the test error on Φ for all five runs when rank $r = 10$ and $r = 20$, respectively. The proposed R-SQN-VR yielded good performance in all runs.



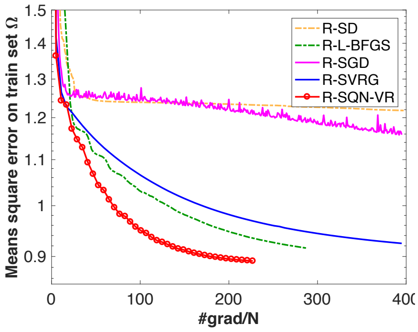
(a-1) run 1



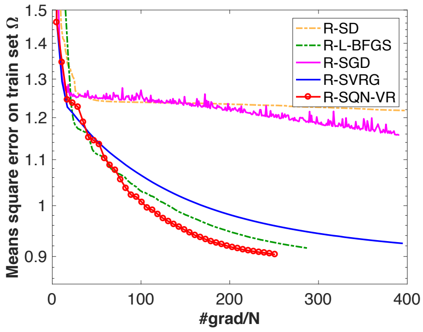
(a-2) run2



(a-3) run 3

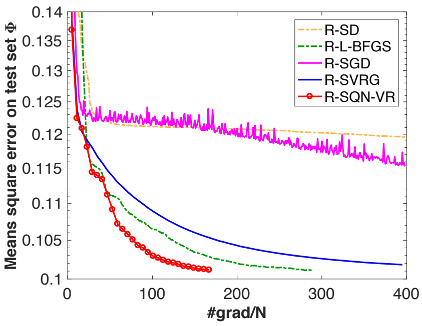


(a-4) run 4

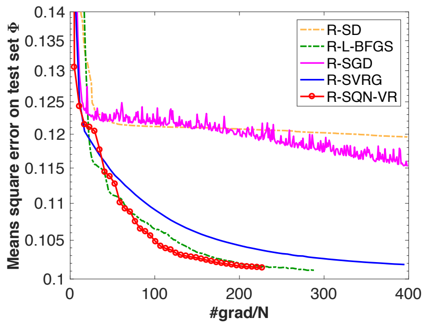


(a-5) run 5

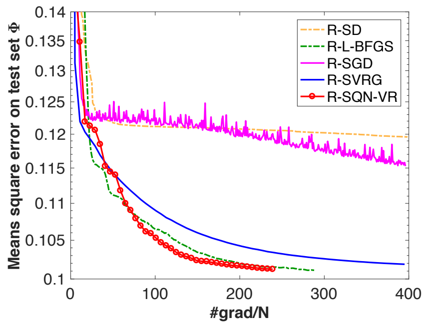
(a) MSE on train set Ω



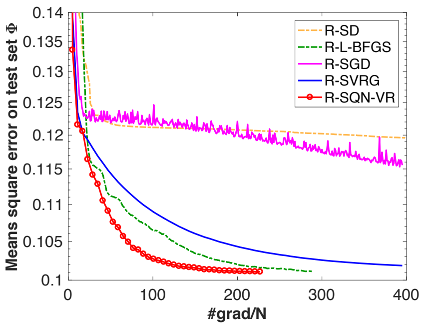
(b-1) run 1



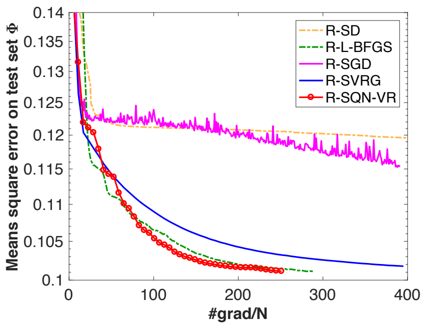
(b-2) run2



(b-3) run 3



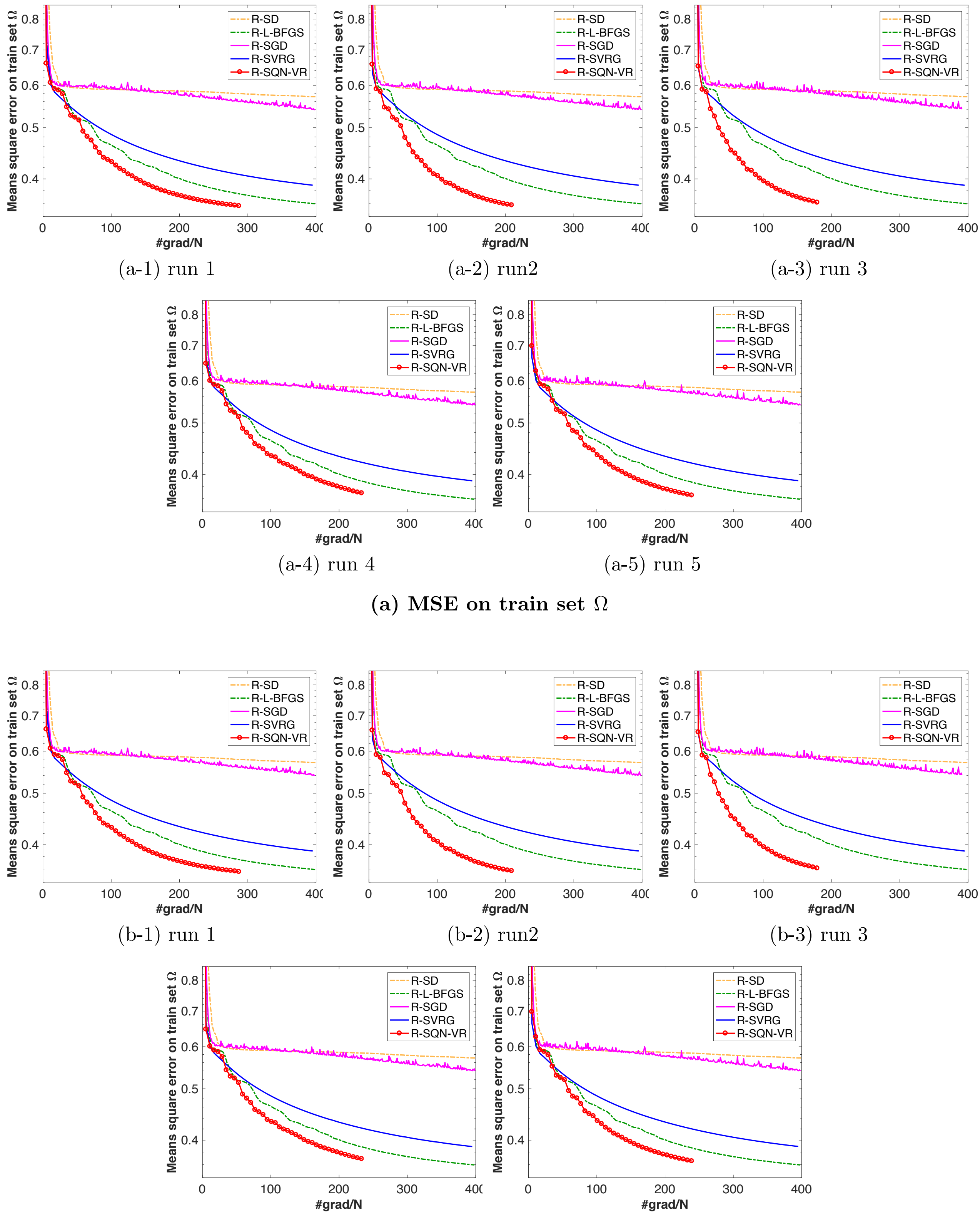
(b-4) run 4



(b-5) run 5

(b) MSE on test set Φ

Figure A.8: Performance evaluations on low-rank MC problem (**MC-R1: lower rank**).



(a-1) run 1 (a-2) run2 (a-3) run 3

(a-4) run 4 (a-5) run 5

(a) MSE on train set Ω

(b-1) run 1 (b-2) run2 (b-3) run 3

(b-4) run 4 (b-5) run 5

(b) MSE on test set Φ

Figure A.9: Performance evaluations on low-rank MC problem (**MC-R2: higher rank**).